Supplementary Information

Flexible Self-rectifying Synapse Array for Energy-efficient Edge Multiplication in Electrocardiogram Diagnosis

Younghyun Lee[1,†], Hakseung Rhee[1,†], Geunyoung Kim[1,†], Woon Hyung Cheong[1], Do Hoon Kim[1], Hanchan Song[1], Sooyeon Narie Kay[1], Jongwon Lee[2,*], and Kyung Min Kim[1,*]

[1]Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

[2]Department of Semiconductor Convergence, Chungnam National University (CNU), 99 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

*Correspondence and requests for materials should be addressed to J. Lee and K. M. Kim. (Email: J. Lee: jwlee80@cnu.ac.kr, K. M. Kim km.kim@kaist.ac.kr)

This **Supplementary Information** contains the following materials:

S1. Electrical characterization of f-MDPE (Supplementary Figures 1-6)

S2. Transformation for hardware-aware training (Supplementary Figures 7-10 and Supplementary Tables 1, 2)

S3. Array operation of f-MDPE for ECG diagnosis (Supplementary Figures 11-13 and Supplementary Table 3)

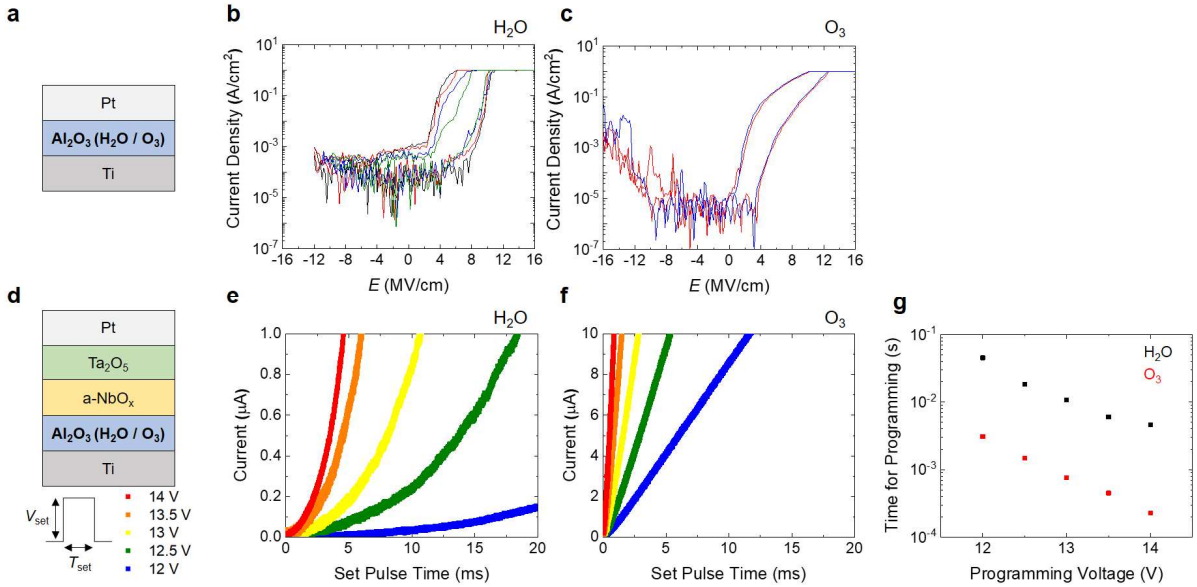
S4. The simulation of a large-scale convolutional neural network (Supplementary Figures 14, 15 and Supplementary Table 4)

S5. Energy consumption and latency comparisons between digital approaches and f-MDPE (Supplementary Figures 16-18 and Supplementary Tables 5, 6)

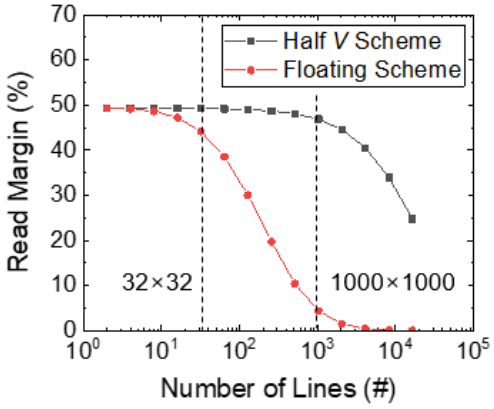
Supplementary Note. Definition of F1 score

Supplementary Video. CMOS-based microprocessor system implanted with a hardware-aware 1D CNN module for individual ECG diagnosis.

S1. Electrical characterization of f-MDPE (Supplementary Figures 1-6)



Supplementary Figure 1 Modifying tunneling layer for enhanced programming speed. **a** A device architecture with a single Al_2O_3 layer and **b,c** *J-E* curves for **b** H_2O-Al_2O_3 and **c** O_3-Al_2O_3 device. Interestingly, both devices exhibited charge trap memristor (CTM) characteristics even though they included only a single Al_2O_3 layer, which is attributed to some charges being trapped within the Al_2O_3 layer. However, the two devices showed a significant difference in switching voltages, with the O_3-Al_2O_3 device switching at a lower voltage. This indicates that charge trapping occurs more easily in the O_3-Al_2O_3 device. We attribute this difference to the lower hydrogen content in the O_3-Al_2O_3 layer. The H_2O-Al_2O_3 inherently contains a higher amount of hydrogen, which passivates trap sites within the amorphous Al_2O_3, hindering the tunneling effect through it. In contrast, O_3-Al_2O_3 reduces the hydrogen content, thereby enhancing the tunneling effect. **d** The CTM device stack, and **e,f** the programming transient current for **e** H_2O-Al_2O_3 and **f** O_3-Al_2O_3-based CTM under various V_{set}. **g** The programming time for each device to program the device to 1 μA by various V_{set}. This enhancement in programming speed will significantly improve the operational speed of the CTM array and reduce the energy required for learning.



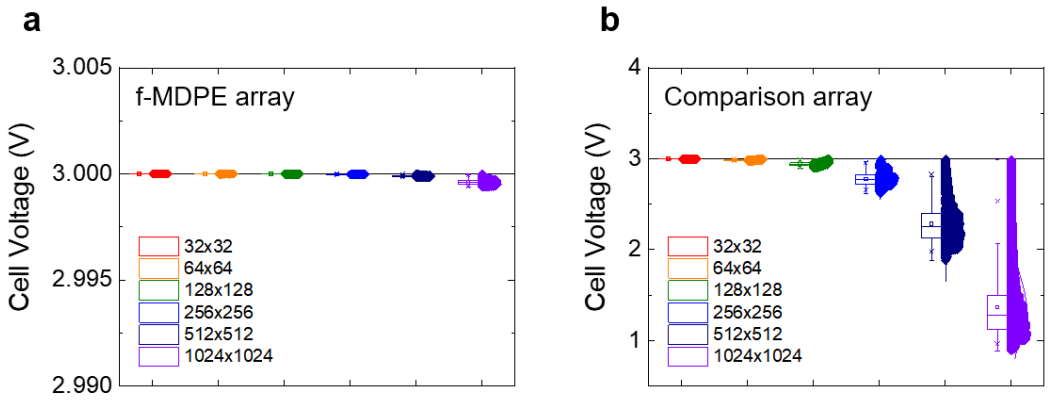
Supplementary Figure 2 Read margin as a function of the number of lines in a square crossbar array. Via a half-voltage scheme, an approximately 50% read margin can be achieved for a 1000×1000 array. The worst-case approximation, assuming that other devices were programmed at the LRS, was used to estimate the maximum array size. The sneak resistance could be assumed to be connected to the selected device in parallel

$$R_{\text{sneak}} = \frac{R_{\text{LRS},1}^{\text{sneak}}}{\text{N}-1} + \frac{R_{\text{LRS},2}^{\text{sneak}}}{(\text{N}-1)^2} + \frac{R_{\text{LRS},1}^{\text{sneak}}}{\text{N}-1}$$

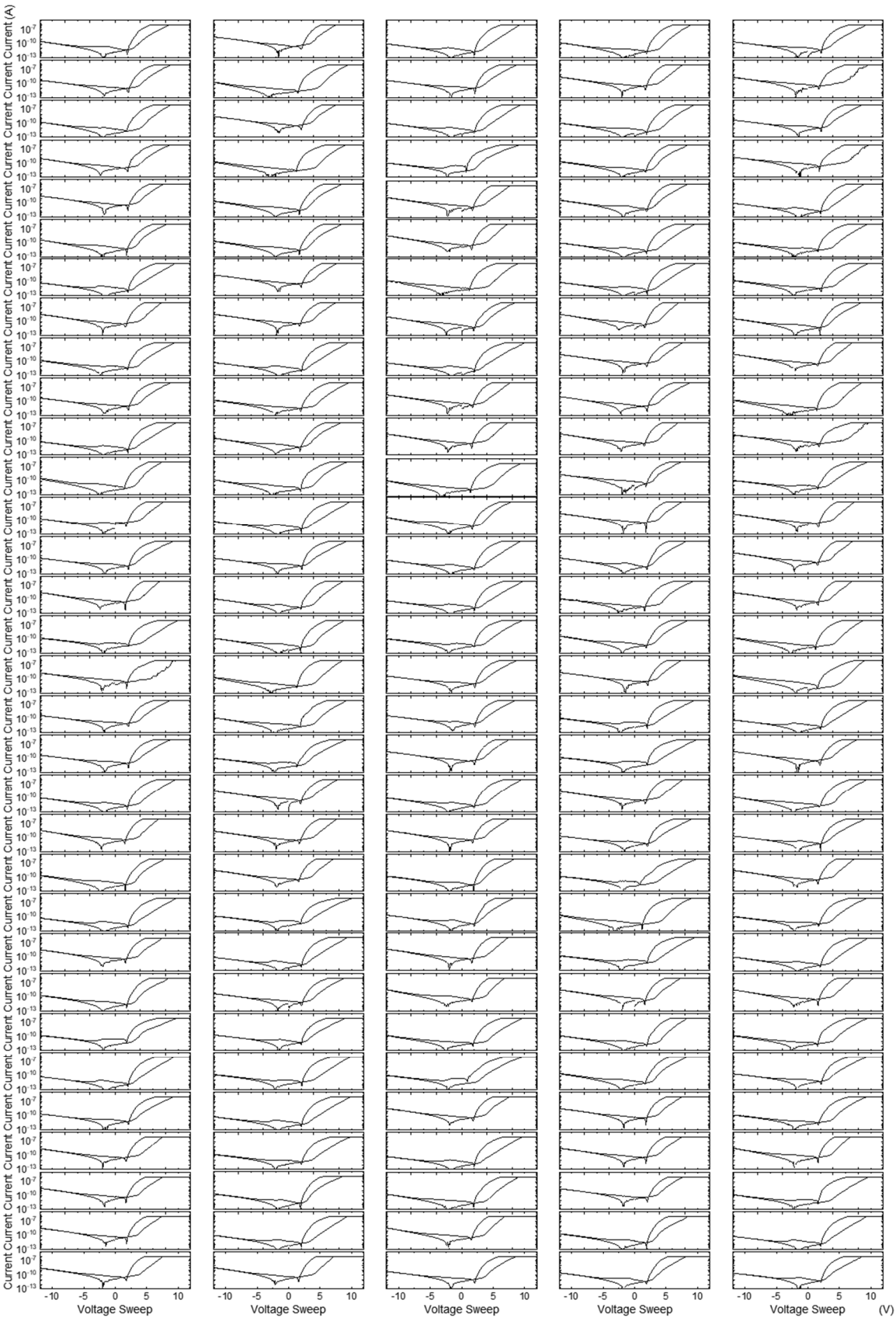
where $R_{\text{LRS},1}^{\text{sneak}}$ represents the resistance of the device that shares the same word line or bit line with the selected device, while $R_{\text{LRS},2}^{\text{sneak}}$ does not. Furthermore, a one-bit pull-up read scheme commonly used for simplified equivalent circuits of an array is considered[1,2]. The pull-up voltage (V_{pu}) is applied to the selected word line, with the pull-up resistance (R_{pu}) connected in series. We calculated V_{out} (the voltage across the selected cell), where V_{pu} is equal to twice the read voltage and R_{pu} is equal to the LRS resistance[3,4]. Consequently, the read margin (RM) can be defined as

$$RM \equiv \frac{\Delta V_{\text{out}}}{V_{\text{pu}}} = \frac{V_{\text{out,HRS}}}{V_{\text{pu}}} - \frac{V_{\text{out,LRS}}}{V_{\text{pu}}}$$

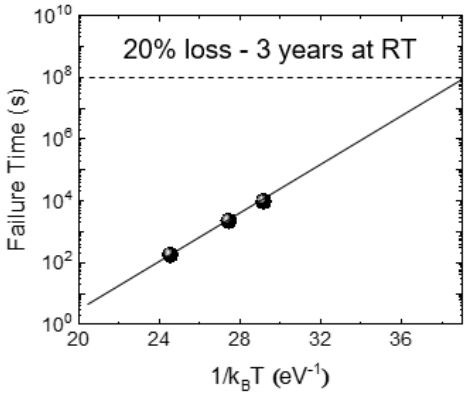
where V_{out} is the voltage across the selected device.



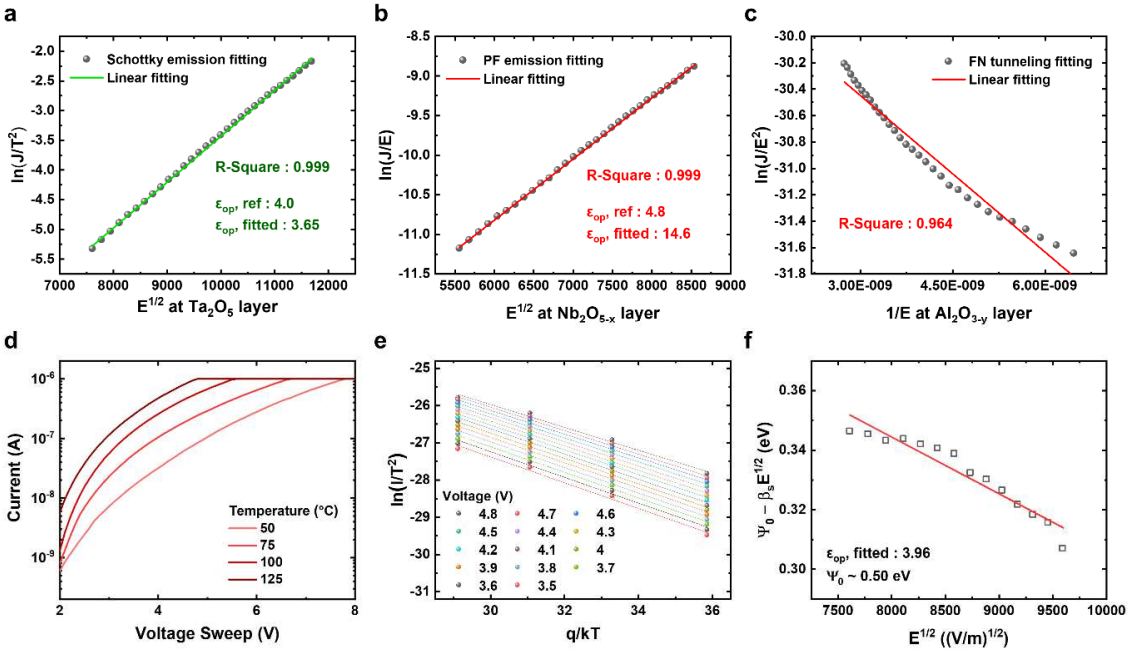
Supplementary Figure 3 Cell voltages considering line resistance with various array size. **a** Cell voltage distribution of a f-MDPE array with V_{read} = 3 V. **b** Cell voltage distribution of a common memristor array that the conductance between 10 μS and 150 μS with V_{read} = 3 V.



Supplementary Figure 4 The *I-V* curves of 160 devices in the f-MDPE array. The voltage was first swept from 0 V to +12 V and back to 0 V, followed by a sweep from 0 V to −12 V and back to 0 V. A compliance current of 1 μA was used during the DC *I-V* measurements.



Supplementary Figure 5 Arrhenius plot of failure time verses various temperatures. The failure time was defined as the moment at which 20% of the original conductance was lost.



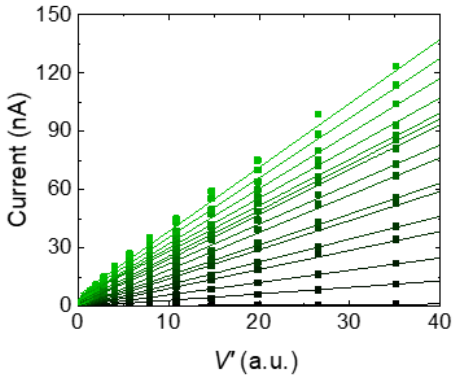
Supplementary Figure 6 Fitting analysis of possible conduction mechanisms in the f-MDPE. Fitting results of **a** Schottky emission ($\ln(J/T^2)$ vs. $E^{1/2}$) at the Ta_2O_5 layer, **b** Poole-Frenkel emission ($\ln(J/E)$ vs. $E^{1/2}$) at the NbO_x layer, and **c** Fowler-Nordheim tunneling ($\ln(J/E^2)$ vs. $1/E$) at the Al_2O_3 layer. For the fitting, the LRS *I-V* curve at room temperature was fitted in the voltage range of +3.5 V to +6.5 V. The local electric field in each layer was calculated using the thickness and dielectric constant of the oxide layers. The thicknesses were 6 nm for Ta_2O_5, 25 nm for NbO_x, and 7 nm for Al_2O_3. The dielectric constants were 24 for Ta_2O_5, 45 for NbO_x, and 9 for Al_2O_3.[5-7] The linear fitting results, along with the reasonable optical dielectric constant (ε_{op}) value, support that the primary conduction mechanism in this device is Schottky conduction at the Ta_2O_5 interface. In case of the Poole-Frenkel fitting, ε_{op} was significantly higher than the reference value. In case of the FN tunneling, the linear fit was not satisfactory. **d** LRS *I-V* curves measured at various temperatures ranging from 50°C to 125°C. **e** Schottky-type plot of $\ln(I/T^2)$ vs. $1/T$ from the current results in **d**. **f** Schottky barrier height at a given electric field ($\Psi_0 - \beta_s E^{\frac{1}{2}}$, where Ψ_0 and β_s represent the zero-field Schottky barrier height and Schottky coefficient, respectively). The extracted dielectric constant of 3.96 closely matched the fitted result in **a**, while the Ψ_0 value was approximately 0.50 eV.

S2. Transformation for hardware-aware training

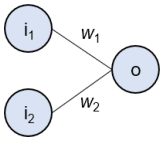
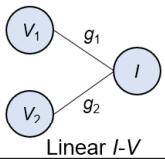
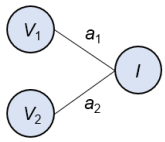
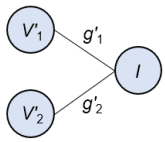
(Supplementary Figures 7-10 and Supplementary Tables 1, 2)

					$k = 0$	$k = 7$
V'	**Input voltage (V)**		b	c	$V' = \exp(-k + b\sqrt{V - c})$	
	3.5		7.66	1.5	5.07E+4	4.62E+1
	3.4				3.85E+4	3.51E+1
	⋮				⋮	⋮
	2.2				6.07E+2	5.54E-1
	2.1				3.77E+2	3.44E-1
	2.0				2.25E+2	2.05E-1
g'	**Analog states**	g_{ref} **(nS)**	a		$g' = \exp(a + k)$ **(×** 10^9**)**	
	16	15	-27.79064		8.52E-4	9.35E-1
	15	14	-27.86691		7.90E-4	8.66E-1
	⋮		⋮			
	3	2	-29.09895		2.30E-4	2.53E-1
	2	1	-30.1967		7.69E-5	8.43E-2
	1	0.1	-32.72548		6.13E-6	6.72E-3

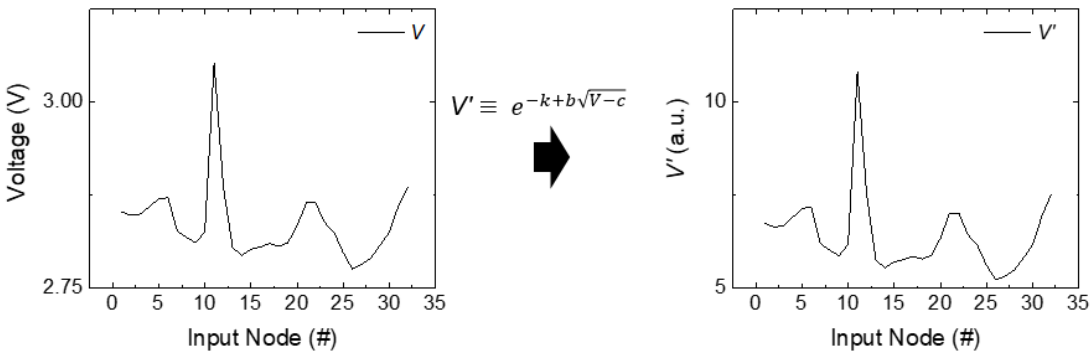
Supplementary Table 1 Comparison of g' and V' for compensation parameter $k = 0$ and $k = 7$ during hardware-aware training. The scale between g' and V' is modified by introducing the k. $k = 0$ can lead to a significant scale mismatch between the g' and V'. Such a mismatch can cause several problems: significant differences in scale can result in unstable gradients during backpropagation, making it difficult to optimize network parameters effectively.



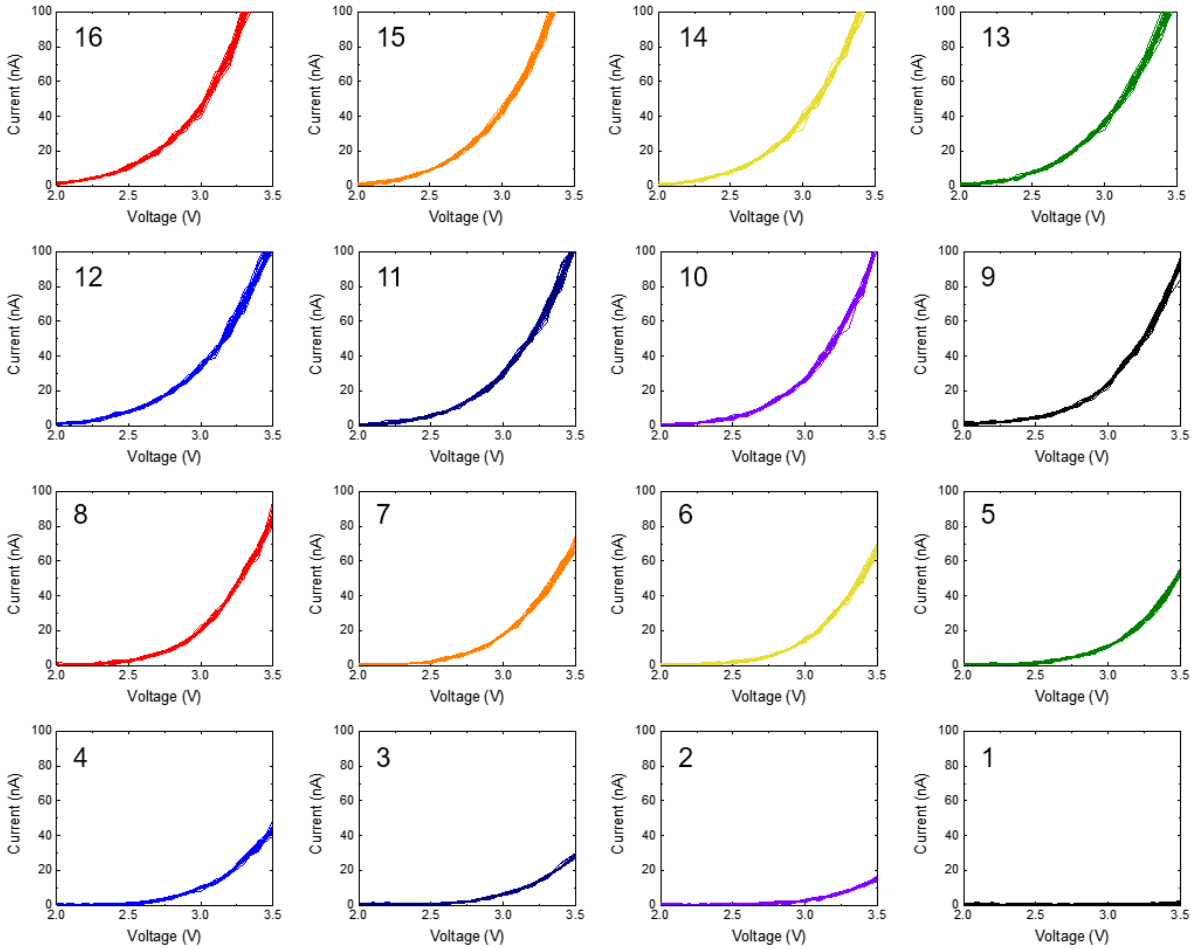
Supplementary Figure 7 Linear behaviors of current vs. transformed voltage (*I*-*V*′ plot) for all conductance states.

	Model	Output (For inference)	Weight Gradient (For training)
Software (Ideal)		$o = w_1 i_1 + w_2 i_2$	$\frac{\partial o}{\partial w_1} = i_1$
Memristor (Ohmic)	 Linear *I-V*	$I = g_1 V_1 + g_2 V_2$	$\frac{\partial I}{\partial g_1} = V_1$
Our Device (Schottky)	 Non-linear *I-V*	$I_{sch}(a,V) = e^{a} e^{b\sqrt{V-c}}$ $I = I_{sch}(a_1, V_1) + I_{sch}(a_2, V_2)$	$\frac{\partial I}{\partial a_1} = e^{a_1} e^{b\sqrt{V_1 - c}}$
	 Transformation	$I = g'_1 V'_1 + g'_2 V'_2$	$\frac{\partial I}{\partial g'_1} = V'_1$

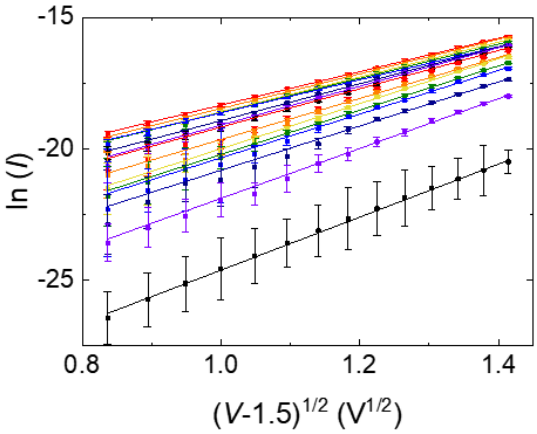
Supplementary Table 2 Comparison of computing a neural network between software (ideal case), memristor (Ohmic conduction), and our device (Schottky conduction). By transformation, the partial derivatives are simplified enabling energy-efficient backpropagation.



Supplementary Figure 8 Transformed input voltage (V') of a common ECG pulse with $k = 7$.



Supplementary Figure 9 *I-V* curves of 16 states at the read voltage region (+2 V ~ +3.5 V). Each curve is obtained from 20 devices, after programming to reach the target conductance at +3 V.



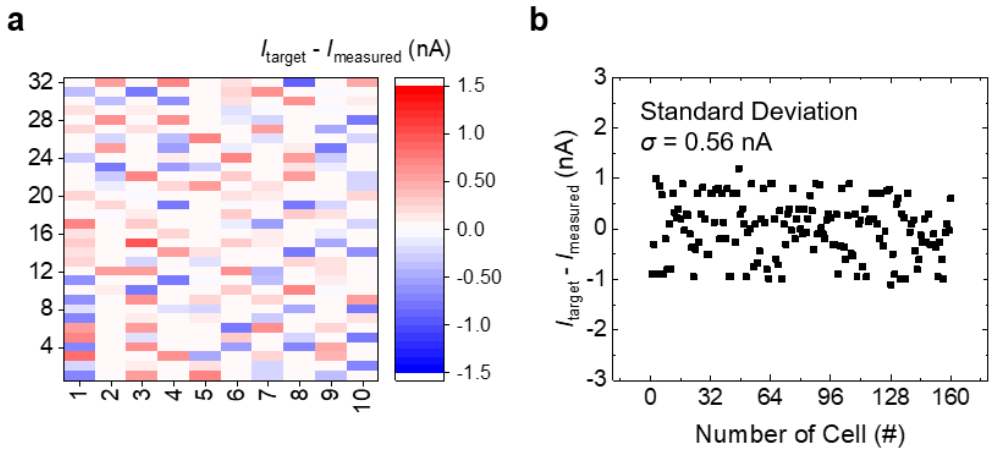
Supplementary Figure 10 Ln(I) vs. $(V\text{-}1.5)^{1/2}$ plots for analog states to the Schottky emission equation corresponding to Supplementary Fig. 9.

S3. Array operation of f-MDPE for ECG diagnosis
(Supplementary Figures 11-13 and Supplementary Table 3)

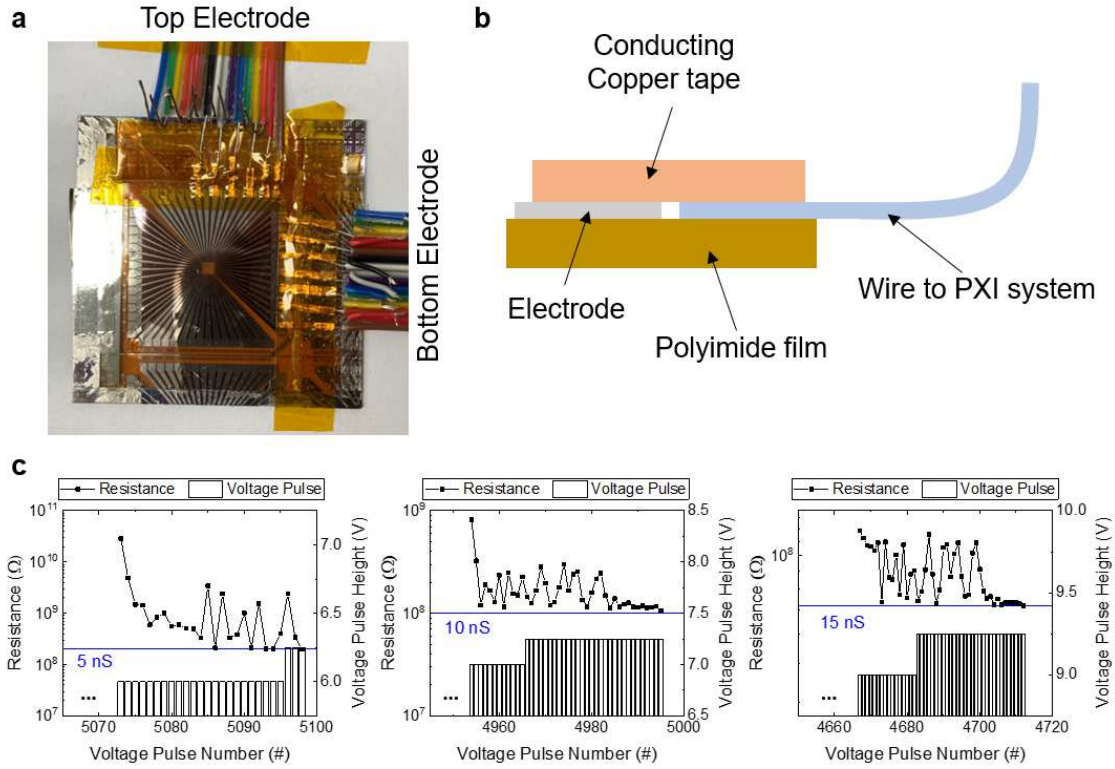
Step	Description
1	Load ECG raw data record.
2	Denoise loaded data using the Mallat algorithm.
3	Load the location and class information for each peak in the annotation file.
4	Extract single pulses by selecting segments from the $(i-100)^{th}$ to the $(i+200)^{th}$ data point for each peak located at position i.
5	Resample each single pulse to a length 32 data by interpolation and normalize from 2.0 V to 3.5 V.

Supplementary Table 3 Preprocessing process for ECG diagnosis with f-MDPE.

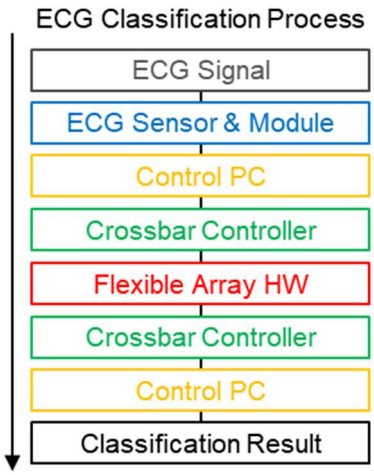
The MIT-BIH dataset is frequently used to evaluate arrhythmia detectors[8]. This dataset offers sequential ECG signals, including peak locations and true labels. Initially, we denoised the raw data using the Mallat algorithm, a well-known method for denoising ECG signals[9]. For each peak located at position i, we extracted single pulse segments ranging from the $(i-100)^{th}$ to the $(i+200)^{th}$ position, creating individual pulse samples for analysis. Subsequently, each pulse was resampled to a length of 32 data points, and the amplitude was normalized from 2.0 V to 3.5 V, corresponding to the input voltage range of the f-MPDE.



Supplementary Figure 11 a Disparity between the target and measured currents depicted in Figures 5c, d. **b** Difference in currents between the target and measured currents at each single point. Unprogrammed cells were excluded due to their low current levels.



Supplementary Figure 12 Demonstration of an array operation with electrical connections between the array and the crossbar controller. **a** An optical image of the array after connection. **b** Schematic of the experimental setup for wiring. **c** Programming results of the device. Voltage pulses were applied until the target conductance states were reached (5 nS, 10 nS, and 15 nS). The voltage pulses were sequentially applied while increasing the height from 4 V to 12 V until they reached the target conductances.

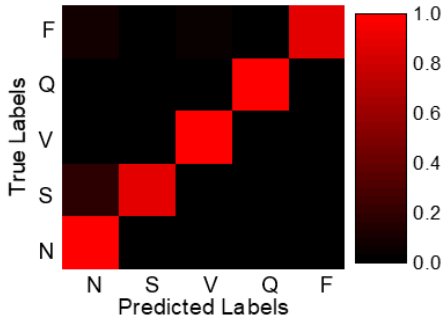


Supplementary Figure 13 A flow chart for real-time ECG diagnosis system as an inference.

S4. The simulation of a large-scale convolutional neural network (Supplementary Figures 14, 15 and Supplementary Table 4)

Array Size		**1k**	**2k**	**4k**
Accuracy	F1 macro	0.8797	0.9372	0.9343
	F1 micro	0.9768	0.9878	0.9884
Conv.1	in	1	1	1
	out	4	4	4
	kernel size	21	21	21
Pool.1	kernel size	9	9	9
Conv.2	in	4	4	4
	out	16	16	16
	kernel size	23	23	23
Pool.2	kernel size	6	6	6
Conv.3	in	16	16	16
	out	32	32	32
	kernel size	25	25	25
Pool.3	kernel size	6	4	3
Conv.4	in		32	32
	out		64	64
	kernel size		27	27
Pool.4	kernel size		4	3
FC1	in	32	128	320
	out	5	5	5
Parameters (#)		**407**	**943**	**1901**
Number of synapses (#)		**814**	**1886**	**3802**

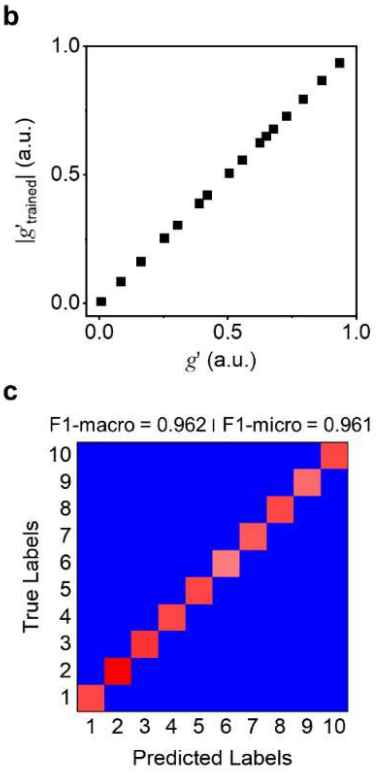
Supplementary Table 4 In the investigation of scaled-up 1D convolutional neural network (1D CNN) model architectures, diverse configurations incorporating array size have been explored. These architectures are designed to control the number of convolutional and pooling layers, as well as fully connected layers, ensuring compatibility with the array size. Notably, the required number of synapses is doubled to accommodate both positive and negative weight values. For example, a 2k-sized array can hold 1000 parameters, which is suitable for operating the proposed 2k model architectures.



Supplementary Figure 14 Confusion matrix of the 2k model.

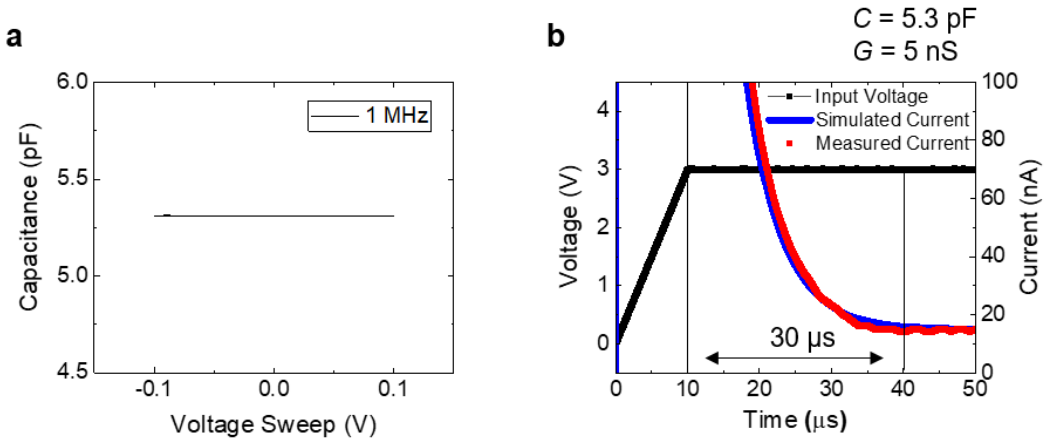
a

Input data	data size	28 x 28
Conv. 1	channels_in	1
	channels_out	4
	kenel size	3 x 3
	stride	1
	padding	1
MaxPool. 1	kernel size	2 x 2
	stride	2
Conv. 2	channels_in	4
	channels_out	16
	kenel size	3 x 3
	stride	1
	padding	1
MaxPool. 2	kernel size	2 x 2
	stride	2
FC	in	784
	out	10

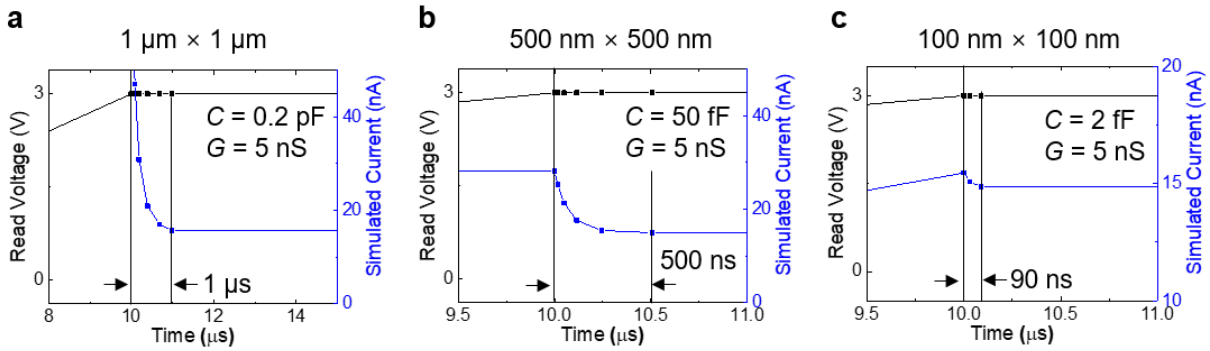


Supplementary Figure 15 a Hardware-aware trained CNN for MNIST classification information. **b** Alignment between trained weights (g'_{trained}) and quantized weights (g'). **c** Confusion matrix of trained MNIST classifier.

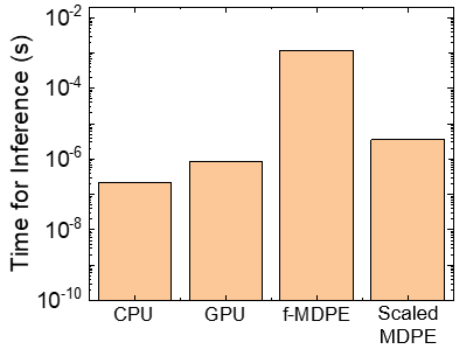
S5. Energy consumption and latency comparisons between digital approaches and f-MDPE (Supplementary Figures 16-18 and Supplementary Tables 5, 6)



Supplementary Figure 16 Read latency of the device. (a) Measured capacitance of the device (5.3 pF). (b) Measured and simulated currents under a read voltage pulse. A numerical simulation was conducted utilizing LTSpice, incorporating a parallel capacitor. The conductance was 5 nS.



Supplementary Figure 17 The simulated read latency of the scaled-down MDPE was investigated for various device sizes; (a) 1 μm × 1 μm, (b) 500 nm × 500 nm, and (c) 100 nm × 100 nm. These simulations were also conducted utilizing LTSpice, incorporating parallel capacitors and considering the reduction in capacitance following the geometric scaling-down effect. Note that we assumed *G* remains unchanged by scaling. This evaluation only considers the effects of latency improvement. In reality, *G* would also decrease with scaling, reducing energy consumption. However, if *G* becomes smaller, it could affect the device's sensing margin and alter its operational conditions. Therefore, it is reasonable to assume that *G* may not change due to improvements in the device, even with scaling.



Supplementary Figure 18 Latency comparisons between f-MDPE and conventional processors while inferring a heartbeat signal. The time to be processed after the last signal input was calculated.

2k model		**Input register**	**MDPE**	**ADC**	**MUX**	**MaxPool**	**Output register**
	Array size	# of operation	# of multiplication	# of conversion	# of operation	# of operation	# of operation
Conv.1	21×4	300	25,200	1,200	1,200	0	300
Pool.1	9×1	300	2700	300	300	0	300
Conv.2	23×4	300	27,600	1,200	1,200	0	300
Pool.2	6×1	288	1,728	288	288	0	288
Conv.3	25×2	288	14,400	576	576	0	288
Pool.3	6×1	192	1152	192	192	0	192
Conv.4	27×2	192	10,368	384	384	0	192
Pool.4	4×1	128	512	128	128	0	128
FC layer	128×5	1	640	5	5	1	1
SUM		**1,989**	**84,300**	**4,273**	**4,273**	**1**	**1,989**

Supplementary Table 5 Operation counts for each component in the 2k model. In the 2k model, input register, analog-to-digital convertor (ADC), multiplexer (MUX), max-pooling (MaxPool), and output register are considered as peripheral circuits. Note that max-pooling is used for final classificating at output neuron in fully connected (FC) layer.

Energy Consumption per Inference	**f-MDPE (5 μm × 5 μm)**		**Scaled-down MDPE (100 nm × 100 nm)**		**CPU**	**GPU**
	Energy (J)	Counts (#)	Energy (J)	Counts (#)		
Input register	1.55×10^{-14}	1,989	1.55×10^{-14}	1,989		
MDPE	1.40×10^{-12}	84,300	4.19×10^{-15}	84,300		
ADC	5.15×10^{-13}	4,273	5.15×10^{-13}	4,273		
MUX	5.71×10^{-15}	4,273	5.71×10^{-15}	4,273		
MaxPool	4.00×10^{-13}	1	4.00×10^{-13}	1		
Output register	1.55×10^{-14}	1,989	1.55×10^{-14}	1,989		
Total	1.20×10^{-7} J		2.64×10^{-9} J		1.74×10^{-4} J	4.24×10^{-5} J

Supplementary Table 6 Energy consumption of the 2k model. The detailed energy consumption of the 2k model was compared between digital approaches and MDPE. For MDPE, the average conductance state is 5.161 nS in 2k model, and we assume that the average input read voltage is 3 V. Then, we calculated the energy consumption of the f-MPDE per multiplication (5 μm × 5 μm) as $E = V^2 \times G \times t = 1.40 \times 10^{-12}$ J, where t is the read latency (t = 30 μs). While the read latency could be reduced under a scaled-down device for 90 ns, as described in Supplementary Fig. 17, the energy consumption per multiplication of 100 nm × 100 nm MDPE was estimated to be 4.19×10^{-15} J. The feature size of the scaled-down device was estimated by using a general E-beam lithography process, which is sufficient for fabricating a 100-nm array. The energy consumption of peripheral circuits was estimated based on 28-nm CMOS technology including I/O register[10], ADC[11], MUX[12], MaxPool[13]. Thus, the total energy consumption for inference was 1.20×10^{-7} J and 2.64×10^{-9} J, respectively. Furthermore, for the digital approaches benchmarked by Su-in Yi et al.[13] and Da Li et al.[14], a CPU was benchmarked based on 22-nm CMOS technology using the Intel Xeon E5-2650 v2, and a GPU was based on 28-nm CMOS technology using the NVIDIA K20. The benchmark CPU and GPU consumed 2,069 pJ and 503 pJ, repectively, per multiplication. The total energy consumption for an inference was 1.74×10^{-4} J and 4.24×10^{-5} J, respectively.

Supplementary Note. Definition of F1 score

The F1 score serves as a crucial evaluation parameter for assessing the performance of classification algorithms, particularly for imbalanced datasets such as the MIT-BIH ECG dataset. In classifying normal signals, the confusion matrix comprises true positive (TP, correctly identified normal signal), true negative (TN, correctly identified abnormal signal), false positive (FP, incorrectly classified normal signal), and false negative (FN, incorrectly classified abnormal signal).

The accuracy, often called the F1-micro score, is a fundamental metric for evaluating the performance of classification models and is defined as the ratio of TP and TN to the summation of all the elements in the confusion matrix:

$$\text{Accuracy (F1-micro score)} = \frac{\text{TP}+\text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

This metric measures the proportion of correctly classified instances among all instances in the dataset, providing an overall assessment of the model's predictive capability. However, in the case of an imbalanced dataset, the accuracy can be significantly influenced by the imbalance in the dataset, leading to biased evaluations.

To address this issue, the F1 score is calculated as the harmonic mean of precision and recall, where precision represents the ratio of TP to the summation of TP and FP, and recall represents the ratio of TP to the sum of TP and FN:

$$\text{F1 score} = \frac{2}{{}^{1}/_{\text{precision}} + {}^{1}/_{\text{recall}}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1 score provides a balanced evaluation considering both recall and precision. Additionally, the F1-macro score, an average of F1 scores for all indices, is the most important parameter for imbalanced datasets containing multiple classes. Achieving a high F1-macro score is often a primary objective because it indicates the model's ability to effectively identify true signals while minimizing false positives and false negatives for all classes. Alternatively, when focusing on classifying only abnormal signals as abnormal, the F1 score for abnormal signals may be the primary goal.

References

1. Sung, S. H. *et al.* Simultaneous emulation of synaptic and intrinsic plasticity using a memristive synapse. *Nature Communication* **13**, 2811 (2022).

2. Huang, C.-H. *et al.* Self-selecting resistive switching scheme using TiO2 nanorod arrays. *Scientific Report* **7**, 2066 (2017).

3. Linn, E., Rosezin, R., Kügeler, C., & Waser, R. Complementary resistive switches for passive nanocrossbar memories. *Nature Materials* **9**, 403 (2010).

4. Ni, R. *et al.* Controlled majority-inverter graph logic with highly nonlinear, self-rectifying memristor. *IEEE Transaction on Electron Devices* **68**, 4897 (2021).

5. H. C. Huang, T. E. Hsieh, *J. Appl. Polym. Sci*. 117, 1960 (2010).

6. N. Fuschillo, B. Lalevic, N. Annamalai, W. Slusark Jr., *J. Non-Cryst. Solids.* 22, 159 (1976)

7. H. Birey, *J. Appl. Phys*. 48, 5209 (1977)

8. Apandi, Z. F. M., Ikeura, R., & Hayakawa, S. Arrhythmia detection using MIT-BIH dataset: A review. *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, 1 (2018).

9. Gang, S., Liang, R., & Yuhan, G. Research on the method of ECG denoising based on improved wavelet packet. *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, 1425 (2015).

10. Yao, P., et al. Fully hardware-implemented memristor convolutional neural network. *Nature* 577, 641-646 (2020).

11. Cao, T. V., Aunet, S., & Ytterdal, T. A 9-bit 50MS/s asynchronous SAR ADC in 28-nm CMOS. *NORCHIP 2012 IEEE*, 1 (2012).

12. Cui, Y., Wu, C., Li, Z., & Shan, W. Energy-Efficient Design of Large Fan-in Dynamic One-hot Multiplexer in 28nm CMOS Technology. *2021 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA)*, 197 (2021).

13. Yi, S.-i., Kendall, J. D., Williams, R. S., & Kumar, S. Activity-difference training of deep neural networks using memristor crossbars. *Nature Electronics* **6**, 45-51 (2023).

14. Li, D., Chen, X., Becchi, M., & Zong, Z. Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, 477 (2016).